ROBUST MULTI-CAMERA TRACKING FROM SCHEMATIC DESCRIPTIONS

Raúl Mohedano and Narciso García

Grupo de Tratamiento de Imágenes, Universidad Politécnica de Madrid, 28040, Madrid, Spain
{rmp,narciso}@gti.ssr.upm.es; www.gti.ssr.upm.es

ABSTRACT

Although monocular 2D tracking has been largely studied in the literature, it suffers from some inherent problems, mainly when handling persistent occlusions, that limit its performance in practical situations. Tracking methods combining observations from multiple cameras seem to solve these problems. However, most multi-camera systems require detailed information from each view, making it impossible their use in real networks with low transmission rate.

In this paper, we present a robust multi-camera 3D tracking method which works on schematic descriptions of the observations performed by each camera of the system, allowing thus its performance in real surveillance networks. It is based on unspecific 2D detection systems working independently in each camera, whose results are smartly combined by means of a Bayesian association method based on geometry and color, allowing the 3D tracking of the objects of the scene with a Particle Filter. The tests performed show the excellent performance of the system, even correcting possible failures of the 2D processing modules.

Index Terms— Multi-camera tracking, 3D tracking, Particle filter, Bayesian association.

1. INTRODUCTION

In the recent years there has been a growing interest for robust monitoring of public and private environments. For that reason, many researchers study how to improve the robustness of visual tracking methods, consisting traditionally in monocular algorithms [1], by using multiple cameras with overlapped fields of view.

Different approaches for multi-camera tracking have been proposed. Some works relate the different views using only planar homographies [2], whereas others use fully calibrated cameras to combine the available multi-camera information [3]. However, independently of the approach, most multi-camera systems gather all the information registered by each camera into a central node, where all the reasonings about coherence between views are performed. Thus, the network should have a sufficient transmission rate to transmit all the information needed at the desired frame rate. So, it would be very interesting to limit as much as possible the amount of information describing each view.

The aim of this work is to perform robust 3D tracking of multiple objects by means of a network of multiple calibrated cameras with moderate transmission and computational capabilities. For that purpose, the proposed 3D system works on simple, schematic descriptions of the objects observed by each camera separately. In that way, mono-camera information could be easily transmitted through the network to the central node responsible for all the 3D reasonings, as they are but simple, light descriptions of the observed situation.

This work has been partially supported by the Ministerio de Ciencia e Innovación of the Spanish Government under project TEC2007-67764 (SmartVision). Also, R. Mohedano wishes to thank the Comunidad de Madrid for a personal research grant.

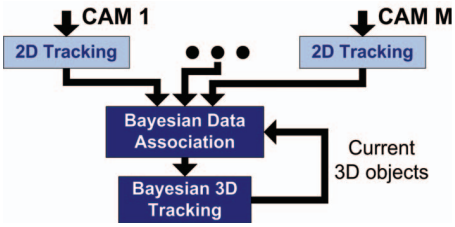

Fig. 1. Block diagram of the proposed system.

2. GENERAL DESCRIPTION OF THE SYSTEM

The 3D reasonings performed in the central node at each time step can be split into two different levels: first, 2D object observations reported by each camera are associated to the currently tracked 3D objects, modeled using Particle Filters [4], by means of a Bayesian approach; and finally, possible entrance of new 3D objects is analyzed, by examining those reported 2D object not associated previously. The resulting 3D information is fed back into the system, so as to use the estimated 3D objects for Bayesian association of 2D observations at the next time step, as depicted in Fig. 1.

As stated previously, the system will work on simple schematic descriptions of the mono-camera observations consisting only of a 'relevant' 2D position and a normalized color histogram per detected object. By 'relevant' 2D position we mean a 2D point which can be considered the projection of a certain 3D point describing the 3D position of an actual 3D object. A possible example of 'relevant' 2D position is the 2D centroid of a detected object, as it could be considered the projection of the centroid of the actual observed 3D object. For people tracking, however, best 'relevant' 2D positions could be proposed: the lowest (central) point of a 2D observed object could be considered the projection of the feet position of an actual person, and the top-most (central) point the projection of the 3D head position as well. Without loss of generality, the experiments presented in this paper have been conducted using the latter, 'head positions', as they have been shown highly stable in common occlusions [5].

As the 'input' of the 3D system has been stated in such a general and simple way, the mono-camera processing performed independently in each camera of the system hardly has restrictions. Therefore, different mono-camera processing approaches could be performed locally in each camera, from simple movement-based object detectors [6], to complex multi-object 2D trackers [7], on the only condition that both 2D position and normalized color histogram for each detected object at each frame are provided.

3. BAYESIAN ASSOCIATION BETWEEN OBSERVATIONS AND 3D OBJECTS

We will estimate the true correspondence between 2D views and currently tracked 3D objects following a Bayesian approach. For that purpose, we should express the probability density of a certain set of views $\mathbf{V}^t$ across the cameras of the system given a set $\mathbf{O}^t$ of 3D objects: that is, $p\big(\mathbf{V}^t|\mathbf{O}^t\big)$. This *pdf* is analyzed in Subsection 3.1. Once this probability (density) has been expressed, we will considered as true the association $\hat{\mathbf{V}}^t$ between reported 2D views and current 3D objects with higher probability (density). Written formally,

$$\hat{\mathbf{V}}^t = \underset{\mathbf{V}^t \in \{\textbf{2D tracks}\}}{\arg\max}\; p\big(\mathbf{V}^t|\mathbf{O}^t\big)\,. \tag{1}$$

As this search implies the exhaustive consideration of every possible combination between reported 2D views and current 3D objects, we will need to establish an efficient algorithm for the search of the best combination. This search is addressed in Subsection 3.2.

3.1. Probability (density) of the 2D observations

First of all, let us set the notation used in the paper. Let us assume that N different 3D objects have been correctly tracked up to time step t using a network composed of M cameras. Let us denote by $\mathbf{v}_{o_n}^{t,c_m}$ the observation of the n-th currently tracked 3D object o_n from camera c_m at time t. For compactness, let us also denote by $\mathbf{V}^t$ the set of all the observations of all the objects at time t, that is, $\mathbf{V}^t = \{\mathbf{V}_{o_1}^t, \ldots, \mathbf{V}_{o_N}^t\}$, where $\mathbf{V}_{o_n}^t = \{\mathbf{v}_{o_n}^{t,c_1}, \ldots, \mathbf{v}_{o_n}^{t,c_M}\}$, and by $\mathbf{O}^t$ the characteristics of all the currently tracked 3D objects, that is, $\mathbf{O}^t = \{o_1^t, \ldots, o_N^t\}$.

3.1.1. Probability of the 2D views given known 3D objects

If the state $\mathbf{O}^t$ of the N actual 3D objects at time step t were known, we could calculate the probability density of a certain set of 2D views $\mathbf{V}^t$ across the cameras of the system given $\mathbf{O}^t$. Assuming conditional independence of the 2D views of each 3D object given the true state of all 3D objects, and also conditional independence of the 2D views from different cameras, we can write

$$p\big(\mathbf{V}^t|\mathbf{O}^t\big) = \prod_{m=1}^{M}\prod_{n=1}^{N} p\big(\mathbf{v}_{o_n}^{t,c_m}|o_n^t\big)\,. \tag{2}$$

This result indicates that the probability density of a proposed combination between reported 2D observations and 3D objects can be conveniently decomposed into a product of simple probabilities, which will be very useful in the search of the best combination.

If we admit that a certain 3D object o_n^t might not be seen from some of the cameras, the product will be limited only to those in which the object o_n^t has one associated view $\mathbf{v}_{o_n}^{t,c_m}$. Thus, omitting one term of this product will be, in practice, equivalent to consider the probability density of the omitted 2D view equal to 1. Therefore, to encourage the association of 'good' 2D views and prevent the 'bad' ones, $p\big(\mathbf{v}_{o_n}^{t,c_m}|o_n^t\big)$ should be defined so as to respect the following consideration: the probability density of a view $\mathbf{v}^{t,c_m}$ that does actually correspond to the 3D object o_n should be clearly greater than 1, whereas a view which does not correspond should have a probability density clearly less than 1.

As for the probability density $p\big(\mathbf{v}_{o_n}^{t,c_m}|o_n^t\big)$, we propose a model made up of two separate factors: one concerning spatial coherence between observed and actual 3D position, and the other considering color coherence, so

$$p\big(\mathbf{v}_{o_n}^{t,c_m}|o_n^t\big) \equiv p\big(C_{o_n}^{t,c_m}|C_{o_n}^t\big)\; p\big(\mathbf{h}_{o_n}^{t,c_m}|\mathbf{h}_{o_n}^t\big)\,. \tag{3}$$

The observed 2D position $\mathbf{h}_{o_n}^{t,c_m}$ has been considered normally distributed, with mean equal to the projection of the 3D position $\mathbf{h}_{o_n}^t$, and covariance proportional to the distance between the object and the camera. The distribution of the Bhattacharyya distance between the observed color histogram $C_{o_n}^{t,c_m}$ and the color histogram $C_{o_n}^t$ of the actual 3D object has been assumed exponential.

3.1.2. Probability of the 2D views given previous views

Eq. (2) assumes that the state $\mathbf{O}^t$ of the N actual 3D objects is known. However, this is not true. As the 3D objects have been tracked using Particle Filters up to time $t-1$, we have a particle-based estimation of the distribution $p\big(\mathbf{O}^{t-1}|\mathbf{V}^{t-1},\ldots,\mathbf{V}^{t_0}\big)$. Thus, instead of $p\big(\mathbf{V}^t|\mathbf{O}^t\big)$, we should use $p\big(\mathbf{V}^t|\mathbf{V}^{t-1},\ldots,\mathbf{V}^{t_0}\big)$ to model the probability of the views $\mathbf{V}^t$.

Assuming that the the 2D views $\mathbf{V}^t$ of the 3D objects $\mathbf{O}^t$ at time t only depend on their true state $\mathbf{O}^t$, we can write

$$\begin{aligned} p\big(\mathbf{V}^t|\mathbf{V}^{t-1},\ldots,\mathbf{V}^{t_0}\big) &= \int p\big(\mathbf{V}^t,\mathbf{O}^t|\mathbf{V}^{t-1},\ldots,\mathbf{V}^{t_0}\big)\,\mathrm{d}\mathbf{O}^t = \\ &= \int p\big(\mathbf{V}^t|\mathbf{O}^t\big)\; p\big(\mathbf{O}^t|\mathbf{V}^{t-1},\ldots,\mathbf{V}^{t_0}\big)\,\mathrm{d}\mathbf{O}^t, \end{aligned} \tag{4}$$

where $p\big(\mathbf{O}^t|\mathbf{V}^{t-1},\ldots,\mathbf{V}^{t_0}\big)$ is the predicted distribution for the state of the 3D objects at time t. This can be expressed as

$$\begin{aligned} &p\big(\mathbf{O}^t|\mathbf{V}^{t-1},\ldots,\mathbf{V}^{t_0}\big) = \\ &\quad = \int p\big(\mathbf{O}^t|\mathbf{O}^{t-1}\big)\, p\big(\mathbf{O}^{t-1}|\mathbf{V}^{t-1},\ldots,\mathbf{V}^{t_0}\big)\,\mathrm{d}\mathbf{O}^{t-1}, \end{aligned} \tag{5}$$

which makes use of the distribution $p\big(\mathbf{O}^{t-1}|\mathbf{V}^{t-1},\ldots,\mathbf{V}^{t_0}\big)$, available as a weighted particle approximation $\{w_{(s)}^{t-1}, \mathbf{O}_{(s)}^{t-1}\}$, $s \in \{1,\ldots,S\}$, where S is the number of particles and $w_{(s)}^{t-1}$ is the weight of the s-th particle $\mathbf{O}_{(s)}^{t-1}$. It is possible to express analytically the predicted distribution in terms of this particle representation as

$$p\big(\mathbf{O}^t|\mathbf{V}^{t-1},\ldots,\mathbf{V}^{t_0}\big) \approx \sum_{s=1}^{S} w_{(s)}^{t-1}\, p\Big(\mathbf{O}^t|\mathbf{O}_{(s)}^{t-1}\Big)\,, \tag{6}$$

which allows the use of Metropolis-Hastings [8] to obtain an equally-weighted sampled version of $p\big(\mathbf{O}^t|\mathbf{V}^{t-1},\ldots,\mathbf{V}^{t_0}\big)$.

Using this equally-weighted sampled version $\{\frac{1}{S}, \mathbf{O}_{(s)}^t\}$ of the predicted distribution, and taking also into account Eq. (2), we could also write Eq. (4) as

$$p\big(\mathbf{V}^t|\mathbf{V}^{t-1},\ldots,\mathbf{V}^{t_0}\big) \approx \frac{1}{S}\sum_{s=1}^{S}\prod_{m=1}^{M}\prod_{n=1}^{N} p\big(\mathbf{v}_{o_n}^{t,c_m}|o_{n,(s)}^t\big)\,. \tag{7}$$

This expression, although very interesting, is not very useful for the association process, as the summation complicates the practical search of the best combination of 2D views to each tracked 3D object. Therefore, for association purposes, instead of using S particles to represent $p\big(\mathbf{O}^t|\mathbf{V}^{t-1},\ldots,\mathbf{V}^{t_0}\big)$, we will content ourselves with a simpler but more practical 'one-point representation' $\hat{\mathbf{O}}^t$ of the predicted distribution. This $\hat{\mathbf{O}}^t$ can be easily estimated from $\{\frac{1}{S}, \mathbf{O}_{(s)}^t\}$ by means of a point-estimator. Using $\hat{\mathbf{O}}^t$, Eq. (4) yields

$$p\left(\mathbf{V}^t|\mathbf{V}^{t-1},\ldots,\mathbf{V}^{t_0}\right) \approx \prod_{m=1}^{M}\prod_{n=1}^{N} p\big(\mathbf{v}_{o_n}^{t,c_m}|\hat{o}_n^t\big)\,, \tag{8}$$

which clearly resembles Eq. (2).

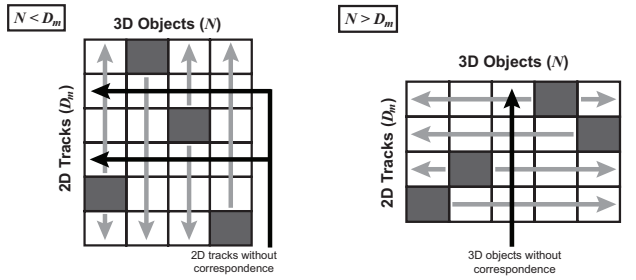

Fig. 2. Situation where the optimal combination can be directly discovered. Left: best 2D view $\mathbf{v}_d^{t,c_m}$ for each 3D object o_n^t marked in dark. Right: best 3D object o_n^t for each 2D view $\mathbf{v}_d^{t,c_m}$ also marked.

3.2. Efficient search of the best association

Once we have defined the probability distribution of the views of the tracked 3D objects at time t, we should find the most probable association of reported 2D views from all cameras and 3D objects. To simplify this search, we will assume that each 3D object o_n^t can only generate (at most) one observed 2D object at each camera, and, additionally, that each observed 2D object can only be produced by one actual 3D object.

One specially interesting consequence of Eq. (8) is the independence of the association performed in each camera: the best global association of 2D views from all cameras and 3D objects will be, at the same time, the best combination of the 2D views of each camera independently. Thus, we will discuss only the search, for each camera separately, of the best one-to-one association of the 2D objects and the currently tracked 3D objects.

Let us suppose that D_m different observed 2D objects $\mathbf{v}_d^{t,c_m}$, $d \in \{1,\dots,D_m\}$, have been reported by camera c_m. We define

$$\bar{\bar{P}} = \{p_{d,n}\}_{\substack{1\le d\le D_m \\ 1\le n\le N}}, \quad \text{where } p_{d,n} = p\big(\mathbf{v}_d^{t,c_m}|\hat{o}_n^t\big). \tag{9}$$

This matrix will form the basis of the proposed search algorithm. The first step of the association algorithm will be the computation of $\bar{\bar{P}}$. This does not imply a great computational cost, as it only requires the computation of $D_m \times N$ simple evaluations of the observation model for each 3D object in the considered camera.

Let us suppose that there are more 2D objects reported by camera c_m than 3D objects ($D_m > N$). Then, each 3D object o_n^t 'would like to choose' the reported 2D object $\mathbf{v}_d^{t,c_m}$ for which its $p\big(\mathbf{v}_d^{t,c_m}|\hat{o}_n^t\big)$ is maximum. So, if each object reaches its maximum probability for a different $\mathbf{v}_d^{t,c_m}$, that is, if

$$d_n = \arg\max_d p\big(\mathbf{v}_d^{t,c_m}|\hat{o}_n^t\big) \tag{10}$$

is different for each 3D object o_n^t, then this particular association yields the optimal combination for the camera, as each o_n^t can reach its ideal associated 2D view. For $D_m < N$, analogously, if each $\mathbf{v}_d^{t,c_m}$ has a different preferred 3D object, that is, if

$$n_d = \arg\max_n p\big(\mathbf{v}_d^{t,c_m}|\hat{o}_n^t\big) \tag{11}$$

is different for each 2D view $\mathbf{v}_d^{t,c_m}$, then the ideal situation will be reachable, and will yield the optimal combination. These two situations, $D_m > N$ and $D_m < N$, in which the optimal combination can be directly discovered are shown in Fig. 2. Unfortunately, the best combination is not usually so clear.

As discussed in Subsection 3.1, considering that o_n^t has no 2D view in camera c_m associated would be, in probability, equivalent to considering a 2D view with a probability density of 1. Therefore, no 3D object o_n^t will ever be associated to a 2D view $\mathbf{v}_d^{t,c_m}$ with probability density less than 1, as considering that it cannot be seen from that camera will always be preferable. Thus, the associations with $p_{d,n} < 1$ will be considered 'invalid', and then:

- If the 2D view $\mathbf{v}_d^{t,c_m}$ is the best possible view for the 3D object o_n^t, and no other 3D object $o_{n'}^t$, $n' \neq n$, considers it valid, then o_n^t will choose $\mathbf{v}_d^{t,c_m}$ or none. Therefore, every $p_{d',n}$ with $d' \neq d$ will be set as invalid.
- Analogously, if the 3D object o_n^t is the most probable one for the 2D view $\mathbf{v}_d^{t,c_m}$, and no other view $\mathbf{v}_{d'}^{t,c_m}$, $d' \neq d$, is valid for o_n^t, then $\mathbf{v}_d^{t,c_m}$ could only be chosen by o_n^t.

Applying iteratively both considerations on the elements of $\bar{\bar{P}}$, it is usually possible to reach one of the final forms described by Eqs. (10) and (11). In some situations, however, the indicated procedure is unable to simplify completely $\bar{\bar{P}}$. In that case, we will fix the association $\mathbf{v}_d^{t,c_m} \leftrightarrow o_n^t$ with higher probability, allowing further simplifications of $\bar{\bar{P}}$.

4. 3D TRACKING USING PARTICLE FILTERS

Once the correspondence between 3D objects and reported 2D objects has been established, we can perform the updating step of the Particle Filter according to the expression

$$p\left(\mathbf{O}^t|\mathbf{V}^t,\dots,\mathbf{V}^{t_0}\right) \propto p\big(\mathbf{V}^t|\mathbf{O}^t\big)\, p\big(\mathbf{O}^t|\mathbf{V}^{t-1},\dots,\mathbf{V}^{t_0}\big). \tag{12}$$

As explained in Subsection 3.1.2, we approximate the predicted distribution $p\big(\mathbf{O}^t|\mathbf{V}^{t-1},\dots,\mathbf{V}^{t_0}\big)$ by a set of equally-weighted particles $\{\frac{1}{S}, \mathbf{O}_{(s)}^t\}$ obtained through Eq. (6) by means of the Metropolis-Hastings algorithm [8]. Thus, the posterior distribution $p\left(\mathbf{O}^t|\mathbf{V}^t,\dots,\mathbf{V}^{t_0}\right)$ will be approximated by the same set of particles, but with weights proportional to

$$w_{(s)}^t \propto p\Big(\mathbf{V}^t|\mathbf{O}^{t,(s)}\Big) = \prod_{m=1}^{M}\prod_{n=1}^{N} p\big(\mathbf{v}_{o_n}^{t,c_m}|o_{n,(s)}^t\big). \tag{13}$$

The final estimation of the state of the N tracked objects will be the mode of the posterior distribution, which will be assumed the particle $\mathbf{O}^{t,(s)}$ with greater weight $w_{(s)}^t$.

5. DETECTION OF NEW 3D OBJECTS

The Bayesian association step may decide that some of the 2D objects reported by the cameras do not correspond to any of the currently tracked 3D objects. We will consider those remaining 2D objects as potential views of new 3D objects entering the scene.

For that purpose, we will check every pair of reported 3D objects from different cameras, and we will analyze the likelihood of both being the projection of an actual 3D object. To do that, we will create a 'candidate 3D object' from the pair, whose color histogram will be the mean of the color histogram of the two 2D objects under consideration, and whose 3D position will be estimated through triangulation (DLT) from the 2D positions of the 2D objects. Thus, each pair of 2D views will be scored with their probability given the 'candidate object' created from them .

We will assume that only one 3D object can enter the scene at each time step. If the probability of the best analyzed pair exceed a certain threshold, then the 'candidate 3D object' created from them will be considered real, and will be added to the Particle Filter into the next time step.

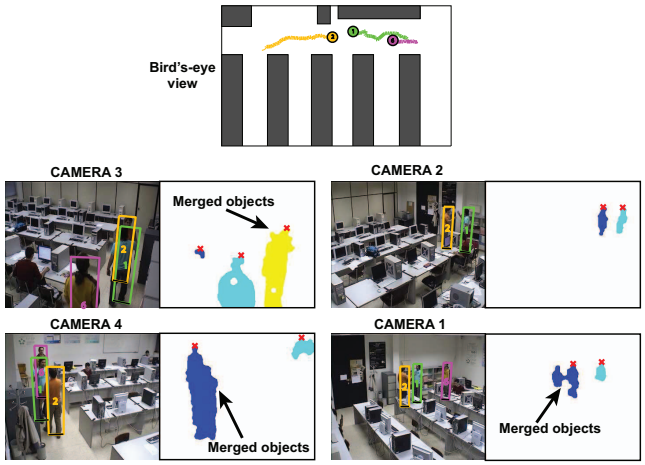

Fig. 3. Tracking results from erroneous 2D object detection.

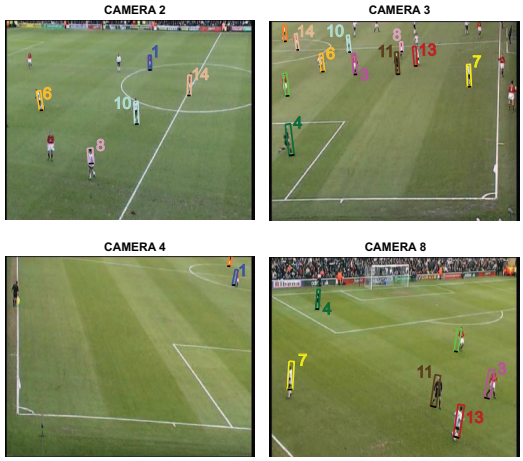

Fig. 4. Tracking results for a complex environment monitored using 8 cameras (only 4 cameras shown).

6. RESULTS

In the presented tests, people tracking has been performed using multiple cameras. For that purpose, head positions have been considered, and the estimated head positions have been used to model each person as a vertical cylinder. The joint state of the tracked 3D objects has been represented using 2000 particles. The dynamic model used does not address interactions between objects, and considers that each object moves with constant velocity. The appearance of the objects has been modeled using normalized color histograms with size 8×8 and considering only the channels r and g (as b is redundant). As for the 2D processing performed in each camera independently, a very simple algorithm for 2D object detection has been used: binary movement detection [6], with subsequent detection of connected blobs. The top-most pixel of each blob has been considered the 2D head position of the object. The histograms has been calculated within each blob. This extremely simple 2D processing proves the capabilities of the system, as produce many erroneous detections when the objects of the scene are interacting closely (see Fig. 3).

Fig. 3 shows the performance of the system on an indoor setting composed of 4 cameras, with 3 people. The figure shows that the system is able to correctly handle severe errors of the 2D detectors. In that controlled situation, both association and detection algorithms show an excellent behavior. Fig. 4 shows the performance of the system on a more complex situation[1], monitored using 8 cameras (although only 4 are displayed here), with more that 20 objects interacting in the scene. It has been observed that the performance of the proposed Bayesian association algorithm is also excellent in this situation, yielding correct associations in short times, and allowing a satisfactory 3D tracking. However, the detection algorithm worsens its performance, even proposing incorrect detections, frequently corrected by the association algorithm after some iterations.

7. CONCLUSION

We have presented an accurate and robust multi-camera 3D tracking method, able to locate and track multiple interacting objects in the 3D world. This method works on simple schematic descriptions of the observations performed by each camera, so it does not imply the transmission of great amounts of information through the communication network connecting the cameras. The simplicity of the descriptions also allows to use a great variety of 2D processing algorithms in each camera, from simple movement detection algorithms to complex multi-object 2D tracking methods.

The system is able to perform a robust 3D tracking from 2D observed objects by analyzing the geometric and color consistence of the reported views. For that purpose, it performs an efficient probabilistic association between reported 2D views and tracked 3D objects, allowing the localization of the objects of the scene over time by means of a Particle Filter. The system has been tested, without loss of generality, for multiple people tracking, showing its capability to successfully address.

[1] SCEPTRE database: *http://sceptre.kingston.ac.uk/sceptre/default.html*

8. REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: a survey," *ACM Computing Surveys (CSUR)*, vol. 38, no. 4, 2006.

[2] S. M. Khan and M. Shah, "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE Trans. PAMI*, vol. 31, no. 3, pp. 505–519, 2009.

[3] A. Mittal and L. S. Davis, "M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene," *Int. Journal of Computer Vision*, vol. 51, no. 3, pp. 189–203, 2003.

[4] O. Cappe, S. J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential monte carlo," *Proceedings of the IEEE*, vol. 95, no. 5, pp. 899–924, 2007.

[5] R. Mohedano, C. R. del-Blanco, F. Jaureguizar, L. Salgado, and N. Garcia, "Robust 3d people tracking and positioning system in a semi-overlapped multi-camera environment," in *Proc. IEEE ICIP*, 2008, pp. 2656–2659.

[6] C. Cuevas, N. García, and L. Salgado, "A new strategy based on adaptive mixture of gaussians for real-time moving objects segmentation," in *SPIE Real-Time Image Proc.*, 2008, vol. 6811.

[7] E. Maggio and A. Cavallaro, "Learning scene context for multiple object tracking," *IEEE Trans. Image Processing*, vol. 18, no. 8, pp. 1873–1884, 2009.

[8] Z. Khan, T. Balch, and F. Dellaert, "Mcmc-based particle filtering for tracking a variable number of interacting targets," *IEEE Trans. PAMI*, vol. 27, no. 11, pp. 1805–1819, 2005.